Our approach is one way to formally bridge a gap between research on emotions in cognitive science, and the formal approaches to rational agent design in AI. Well-defined notion of emotional states is useful for intelligent systems that are to operate under time pressure, in multi-agent environments. First, emotions can serve as control mechanisms that allow agents to manage their computational resources while deliberating about action under time pressure. Second, well defined notions of emotions serve as vocabulary that the agents can use to describe their internal states to each other without referring to implementational details. Finally, these notions are critical when the agents are to effectively interact with humans.

The approach we outlined serves as a point of departure for much of the needed future work. The definitions of emotional transformations can be elaborated upon, and more intuitive special cases can be arrived at. These cases should ultimately find their way into the taxonomy depicted in Figure 1, and be defined in terms of measurable attributes [24]. Further, the dynamic models of emotional states, like the one in Figure 2, can become far more elaborate, thus allowing the agents to predict the emotional states of other agents and humans in much more detail. Our current work involves implementation and experimental validation in simulated air defense environment in which agents act under time pressure and interact with other agents and humans. He expect to show how emotions convey an advantage to rational agents allowing them to act and interact effectively.

References

[1] Robert Axelrod. *The Evolution of Cooperation.* Basic Books, 1984.

[2] Cristina Bicchieri. *Rationality and Coordination.* Cambridge University Press, 1993.

[3] Craig Boutilier, Thomas Dean, and Steve Hanks. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial intelligence Research*, 11:1–94, 1999.

[4] David Carmel and Shaul Markovitch. Learning models of intelligent agents. In *Proceedings of the National Conference on Artificial Intelligence*, pages 62–67, Portland, OR, August 1996.

[5] P. R. Cohen and H. J. Levesque. Rational interaction as the basis for communication. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication.* MIT Press, 1990.

[6] G. Cottrell and Metcalfe. Empath: Face, emotion and gender recognition using holons. In *Advances in Neural Information Processing.* Morgan Kaufman Publishers, 1991.

[7] Tuomas Sandholmand R. H. Crites. Multiagent reinforcement learning and iterated prisoner's dilemma. *Biosystems Journal*, 37:147–166, 1995.

[8] D. Dennett. Intentional systems. In D. Dennett, editor, *Brainstorms.* MIT Press, 1986.

[9] Antonio R. Dimasio. *Descartes' Error.* Grosset/Putnam, 1994.

[10] Jon Doyle. Rationality and its role in reasoning. *Computational Intelligence*, 8:376–409, 1992.

[11] Edmund H. Durfee, Jaeho Lee, and Piotr Gmytrasiewicz. Overeager rationality and mixed strategy equilibria. In *Proceedings of the National Conference on Artificial Intelligence*, July 1993.

[12] N. H. Fridja. *The Emotions.* Cambridge University Press, 1986.

[13] Piotr J. Gmytrasiewicz and Edmund H. Durfee. A rigorous, operational formalization of recursive modeling. In *Proceedings of the First International Conference on Multiagent Systems, ICMAS'95*, pages 125–132, July 1995.

[14] B. Hayes-Roth, B. Ball, C. Lisetti, and R. Picard. Panel on affect and emotion in the user interface. In *Proceedings of the 1998 International Conference on Intelligent User Interfaces*, pages 91–94, 1998.

[15] W. James. What is an Emotion? *Mind*, 9:188–205, 1884.

[16] W. James. The Physical Basis of Emotion. *Psychological Review*, 1:516–529, 1894.

[17] P. N. Johnson-Laird and K. Oatley. Basic Emotions, Rationality, and Folk Theory. *Cognition and Emotion*, 6(3/4):201–223, 1992.

[18] S. Kraus and K. Sycara. Argumentation in negotiation: A formal model and implementation. *Artificial Intelligence*, 104(1-2):1–69, 1989.

[19] Victor Lesser, Michael Atighetchi, Brett Benyo, Raja Bryan Horling, Vincent Anita, Wagner Regis, Ping Thomas, Shelley Xuan, and ZQ Zhang. The intelligent home testbed. In *Proceedings of the Autonomy Control Software Workshop (Autonomous Agent Workshop)*, 1999.

[20] H. Leventhal and K. Scherer. The relationship of emotion to cognition:a functional approach to semantic controversy. *Cognition and Emotion*, 1(1):3 – 28, 1987.

[21] Christine Lisetti. Facial expression recognition using a neural network. In *Proceedings of the Florida Artificial Intelligence Research Symposium*, 1998.

[22] Christine Lisetti and Diane J. Schiano. Automaic facial expression interpretation: Where human interaction, artificial intelligence and cognitive science intersect. *Pragmatics and Cognition, Special Issue on Facial Information Precessing and Multidisciplinary Perpective*, 1999.

[23] M. L. Littman. Markov games as a frameowrk for multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 1994.

[24] Andrew Ortony, Gerald Clore, and Allen Collins. *Cognitive Structure of Emotions.* Cambridge University Press, 1988.

[25] R. Picard. *Affective Computing.* MIT Press, 1997.

[26] Jeffrey S. Rosenschein and Gilad Zlotkin. *Rules of Encounter.* MIT Press, 1994.

[27] Ariel Rubinstein. Finite automata play the repeated prisoner's dilemma. *Journal of Economic Theory*, 39:83–96, 1986.

[28] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall, 1995.

[29] H. Simon. Motivational and Emotional Controls of Cognition. *Psychological Review*, 74:29–39, 1967.

[30] Herbert A. Simon. *Rational Choice and the Structure of the Environment.* MIT Press, 1958.

[31] A. Sloman. Motives, Mechanisms, and Emotions. In M. Boden, editor, *The Philosophy of Artificial Intelligence.* Oxford: Oxford University Press, 1990.

[32] A. Sloman and M. Croucher. Why robots will have emotions. In *Proceedings of the Seventh IJCAI Vancouver, B.C.*, pages 197–202. San Mateo, CA: Morgan-Kaufmann, 1981.

[33] K. Sycara. Multiagent systems. *AI Magazine*, 10(2):79–93, 1998.

[34] Michael Wooldridge and Editors Anand Rao. *Foundations of Rational Agency.* Kluwer Academin Publishers, 1999.

an emotional state. Further, let **IN** be the set of all environmental inputs.

Definition 2: Emotional transformation is a function $EmotTrans : \mathbf{D} \times \mathbf{IN}^* \rightarrow \mathbf{D}$.

Thus, given an initial emotional state, D, and a, possibly empty, history of environmental inputs IN, the value of the $EmotTrans$ function is $EmotTrans(D, IN) = D'$, where D' is the agent's new emotional state. Examples of such emotional transformations are depicted in Figure 2. D' may differ from D in a number of ways. Below we look at some possibilities that correspond to some of the more intuitive emotional states.

Transformations of the action space A

Transformation of the action space A, for example by narrowing the set of alternative actions considered to encompass only a small subset of all of the actions, predisposes the agent to take action from this smaller set. This constitutes the *action tendency* that the emotion is invoking in the agent, as postulated, for example, by Fridja in [12]. In the extreme, narrowing the set A to a single action implements a behavioral condition-response rule.

Formally,
these are transformations $EmotTrans(D, IN) = D'$ such that $D =< P_c(S), A, Proj, U >$, and $D' =< P_c(S), A', Proj, U >$. An emotional transformation that implements an action tendency in one for which $A' \subset A$. For example, an agent becoming angry may result in it considering only a subset of its behavioral alternatives, say, ones of aggressive nature. A special case of this emotional transformation obtains when A' is a singleton set, containing only one behavior. This is an implementation of a emotional condition-action rule; all three emotional states that correspond to the Tit-for-two-Tats strategy in repeated Prisoner's Dilemma game in Figure 2 are of this kind since they result in the agent's being capable of performing only a single behavior.

Another intuitive special case of such transformation is one that results in the agent's deliberating in a more short-term fashion, such as it being rushed or panicked under time pressure. Formally we have: $\forall a_i^{'} \in A' : t_{a_i^{'}} \leq t_{a_i}$, which states that the time horizon of alternative plans considered has diminished. This is characteristic of human decision-makers; people frequently become more short-sighted when they are rushed or panicked, since they have no time to consider long-term effects of their alternative behaviors.

Transformations of the utility functions U

Intuition behind this transformation is that emotions and feelings both implement U, as well as modify it. Humans evaluate desirability of states by having positive or negative feelings about them. Positive or negative emotions or moods may alter these evaluations by, say, decreasing them, as in melancholic or depressed moods (when everything looks bleak), or increasing them, as in elated or happy moods. Other emotional states can change the weights of the factors contributing to the utility ratings (Equation 2).

Formally,
these are transformations $EmotTrans(D, IN) = D'$ such that $D =< P_c(S), A, Proj, U >$, and $D' =< P_c(S), A, Proj, U' >$.

The special case of sadness or melancholy result in evaluation of the desirability of every state to diminish: $\forall s \in S : U'(s) \leq U(s)$.

Transformations of the probabilities of states

The intuition behind this transformation is that changing these probabilities, for instance by simplifying them, can be helpful and save time under time pressure. The most radical simplification is one that makes the most likely state to be the only possible state or result. This corresponds to considering only the most likely result of action and neglecting all less likely states and is often observed in human decision-makers.

Formally, these are transformations $EmotTrans(D, IN) = D'$ such that $D =< P_c(S), A, Proj, U >$, $D' =< P_c^{'}(S), A, Proj^{'}, U >$. The special case described above obtains is when the probability distribution $P_c^{'}$, as well as every projected distribution $P_i^{'}$ returned by the projection function $Proj^{'}$ are deterministic.

Conclusions and Future Work

This paper outlined an approach to formally defining the notions of emotions of rational agents. Following one of the recent approaches to designing rational agents based on decision theory [28], we attempted to define emotional transformations, and the resulting emotional states, as resulting from input the agents receive from the environment. The emotional states are identified as possible modifications of the decision-making situations the agents find themselves in.

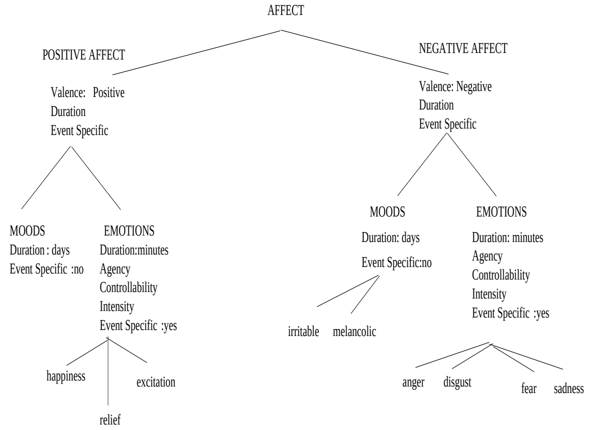

Figure 1: An Example Taxonomy of Emotional States

The taxonomy [24] of emotional states in Figure 1 is aimed at differentiating among emotional states by using values of well-defined attributes. It is clearly desirable to be able to measure the values of the attributes to be able to determine the current emotional state of another agent. Given the emotional state, its impact on decision-making can be modeled. Presently, only some of the attributes differentiating various emotional states are directly measurable; in human agents for example, the positive and negative values of valence attribute may be measurable from facial expression recognition tools. Further, the taxonomy is incomplete in that we (and our colleagues in cognitive science) do not yet know what attributes and their values should be used to differentiate among some emotional states.

Our framework also uses a dynamic model of user's emotional states. Its purpose is twofold: First it is frequently useful not only to assess the current emotional state the other agent is in but also to predict what emotional state will obtain, given the current emotion on one hand, and a system's response or environmental event on the other hand. Second, since some of the variables that determine the current emotional state are not measurable, it may be possible to infer the current state from the previous state, if known, and the environmental input. In Figure 2 we present a very simple example model of emotions' dynamics. It contains only three emotional states; COOPERATIVE, SLIGHTLY ANNOYED, and ANGRY. The transitions among the states are caused by environmental inputs or responses of the system, and they are divided into categories of Cooperative and Uncooperative. Using this dynamic model one can predict that an agent that is in COOPERATIVE emotional state will become SLIGHTLY ANNOYED given Uncooperative input. Further, the emotional state of SLIGHTLY ANNOYED will evolve into ANGRY if another Uncooperative response follows.

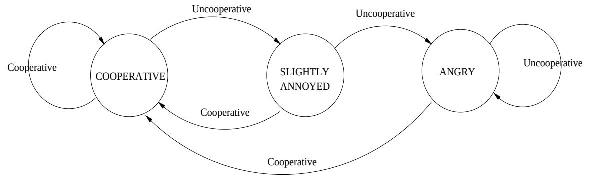

Figure 2: Simple Dynamic Model of an Agent's Emotional States

Dynamic models that use the formalism of finite state automata like the one in Figure 2 are common and used in the field of game theory. The simple example we present here coincides with the Tit-for-two-Tat strategy used for the Prisoner's Dilemma game [1, 4, 27]. The advantage of modeling emotional states of the user with the finite automata model is that models of this kind can be learned, for example using an unsupervised US-L learning algorithm [4, 23, 7].

Decision-Theoretic Definitions of Emotions

We now outline some classes of transformations of the decision-making situation of an agent. We call them *emotional transformations*, and their results are *emotional states*. In other words, an emotional transformation changes one decision-making situation, say a NEUTRAL emotional state, into another one, say an ANGRY emotional state. We assume that the emotional transformations themselves are triggered by some environmental input, IN, the agent experiences. We should caution that our identifying emotions with such transformations does not account for all of the richness of emotions in humans; in fact our decision-theoretic approach limits our formalization to emotions that impact the agent's decision making – the emotions that do not have such impact clearly cannot be accounted for.

Let us denote as $\mathbf{D}$ the set of all decision situations, D, as defined by Definition 1; with the caveat above, we postulate that each $D \in \mathbf{D}$ correspond to they are. Thus $P_c(S)$ fully describes the information the agent has about the present state of the world.

The agent can ponder the consequences of its alternative actions. Due to possible nondeterminism each action, $a_i \in A$, may lead to many resulting possible states. The likelihoods of the resulting states can be specified by another probability distribution, $P_i(S)(\in \mathbf{P})$, also over S. The process of determining the probabilities of different results, i.e., the distribution P_i has been called a probabilistic temporal projection. The projection is a function $Proj : \mathbf{P}(S) \times A \rightarrow \mathbf{P}(S)$; the result of projecting the results of action a_i given the current information about the state $P_c(S)$ results in the projected information about the resulting state, $P_i(S)$: $Proj(P_c(S), a_i) = P_i(S)$. The above formulation does not preclude that the state change due to actions of other agents or exogenous events; here these effects are implicit and folded into the projection function [3].

The desirabilities of the states of the world to the agent are encoded using a utility function $U : S \rightarrow R$, which maps states of the world to real numbers. Intuitively, the higher the utility value of a state the more desirable this state is to the agent. The agent decision problem involves choosing which of the alternative actions in the set A it should execute. One of the central theorems of decision theory states that if the agent's utility function is properly formed, and the agent expresses its uncertain beliefs using probabilities, then the agent should execute an action, a^*, that maximizes the expected utility of the result.

$$a^* = ArgMax_{a_i \in A} \sum_{s \in S} p_i^j U(s^j), \qquad (1)$$

where the p_i^j in the probability the projected distribution $P_i(S)$ assigns to a state $s^j \in S$. Frequently, it is convenient to represent the utility function, U, as depending on a small number of attributes of the states of the world, as opposed to depending on the states themselves. This is intuitive; humans may prefer, say, all of the states in which they have more money, are more famous, and are healthier. The attributes, say wealth, fame, and health are then convenient factors in terms of which the utility function can be expressed. Multi-attribute utility theory postulates that, in some simple cases, the utility of a state is a weighted sum of the utilities, $U(X_l(s))$ of individual attributes:

$$U(s) = \sum_{X_l \in Attributes} W_{X_l} U(X_l(s)), \qquad (2)$$

where the W_{X_l} is the weight, or intuitively, the importance, of the attribute X_l. Having the weights of the attributes explicitly represented is convenient since it enables the tradeoffs among the attributes the agent may have to make. For example, the agent may have to give up some of its wealth to improve its health, and so on.

The elements defined above are sufficient to formally define a decision-making situation of an agent:

Definition 1: A decision-making situation of an agent is a quadruple: $D =< P_c(S), A, Proj, U >$, where S, $P_c(S)$, A, $Proj$ and U are as defined above.

The above quadruple fully specifies the agent's knowledge about the environment, the agent's assessment as to its possible courses of action, the possible results of the actions, and desirability of these results. Our definition here is closely related to that of stochastic processes, and in particular to Markov decision process (see [3, 28] and references therein), but it makes explicit the decision problem the agent is facing by enumerating the alternative action sequences the agent is choosing among.

Given its decision-making situation, and agent can compute its best action, a^*, as specified in Equation 1. It is clear that this computation can be fairly complex. In a multi-agent environment, for example, all of the information the agent has about the physical environment and about the other agents could be relevant and impact the expected utilities of alternative courses of action. Sometimes the agent may have information about the other agents' state of knowledge, which is also potentially relevant. Given these complexities it is clear that a mechanism for managing the agent's computational resources is needed. Here, we suggest that emotional states, as defined below, may provide for such ability.

Emotional States: Classification and Dynamics

As we mentioned, we will view emotions as transformations of the decision-making situation defined above. First, we briefly describe a taxonomy of emotional states and a finite state machine model of dynamics of emotions which can assist agents in measuring and predicting the emotional state of other agents.

of emotions serve as semantics of emotional terms, with which the agents can express their own internal states, understand the states the other agents are in, and thus predict their actions and interact more efficiently. Finally, well defined emotional states of self and others are crucial in the agent's interaction with humans. Frequently, human-computer interaction is impeded by the machine being hopelessly out-of-step with the emotional state of the human user. However, the users' emotional state, such as anger, fear, boredom, panic, surprise, joy, or excitation, can be assessed using measurable and inferred factors (facial expression recognition, vocal intonation, prosody, galvanic skin response, heart rate and breathing patterns, haptic and tactile feedback, body posture [6, 14, 21, 22, 25]), and predicted from dynamic emotion models based on observed input events. For example, it should be possible to predict that an already annoyed human user will not be calmed down by another system response that is not along the user's wishes. Thus, it is important for the machine to model the effect the user's emotional state has on the user's decision-making and his/her tendency for action.

Our approach complements and builds on the existing approaches to designing rational and socially competent agents [2, 3, 5, 10, 11, 13, 18, 26, 28, 33, 34]. Such agents should be able to function efficiently under time and other environmental pressures, and be able to interact and communicate with other agents. This includes informing each other about details of the external environment and about the agents' own internal states, as well as the ability to model and predict the internal states of other agents. Apart from the area of multi-agent systems, our approach has applications in Human-Computer Interaction (HCI) that range from intelligent tutoring systems and distance learning support systems (with recognition of expressions signaling interest, boredom, confusion), to stress and lie detectors, to monitors of pilots and drivers' state of alertness, to software product support systems (with recognition of users being dis/pleased with software products), to entertainment and computer games (enjoyment, confusion), to ubiquitous computing and smart houses [19].

Decision-Theoretic Preliminaries

The objective of our research is to develop a fundamental understanding of the role and usefulness of the concept of emotional states in designing intelligent artificial systems. Our approach draws on and combines an emerging technology of rational agent design of Artificial Intelligence on the one hand [3, 8, 10, 28, 30, 34], with research on human emotions in cognitive science and psychology on the other hand [9, 15, 16, 17, 20, 24, 29, 32, 31].

We use decision-theoretic paradigm of rationality, according to which rational agent should behave so as to maximize the expected utility of its actions (see [3, 10, 28] and references therein). The expected utilities of the alternative courses of action are computed based on their possible consequences, the desirability of these consequences to the agent,[1] and the probabilities with which these consequences are thought by the agent to obtain. The main thrust of our work is to examine ways in which components of the decision-theoretic model, i.e., the utility functions, the set of behavioral alternatives, and the probabilities of consequences, can be transformed in ways that that has been recognized in cognitive science as interactions between emotional states and decision-making.

A rational agent formulates its decision making situation in terms of a finite set, A, of the alternative courses of action, or behaviors, it can execute, which we will call the agent's *action space*. An alternative behavior, say a_i, is a plan consisting of consecutive actions extending into the future time t_{a_i}, which we will call the time horizon of this particular plan. Alternative courses of action in set A can stand for abstract actions as well as for detailed elaborations; increasing the level of abstraction facilitates keeping the size of A down to manageable proportions. We demand that the actions be distinct and that the set A be exhaustive, i.e., that all of the possible behaviors be accounted for. Sometimes an "all-else" behavioral alternative is used for compactness, and represents all other possible behaviors except the ones explicitly enumerated.

At any point, an agent finds itself in some state of the world, but due to the fact that the environment may not be fully observable the agent may be uncertain about the state. The fact that the actual state may be unknown to the agent can be formalized by specifying the set of all possible states of the world, S, together with a family of probability distributions, $\mathbf{P}(S)$, over these states. One of these distributions, say $P_c(S)(\in \mathbf{P})$, specifies which of these states are currently possible and how likely

[1] Such agents are sometimes called self-interested.

Using Decision Theory to Formalize Emotions

Abstract

We use the formalism of decision theory to develop principled definitions of emotional states of a rational agent. We postulate that these notions are useful for rational agent design. First, they can serve as internal states controlling the allocation of computations and time devoted to cognitive tasks under external pressures. Second, they provide a well defined implementation-independent vocabulary the agents can use to communicate their internal states to each other. Finally, they are essential during interactions with human agents in open multi-agent environments. Using decision theory to formalize the notions of emotions provides a formal bridge between the rich bodies of work in cognitive science, and the high-end AI architectures for designing rational artificial agents.

Introduction

Our research is predicated on the thesis that concept of emotions can be formalized and be made useful in designing artificial agents that are to flexibly and adaptively interact with other agents and humans in open multi-agent environments. Our formalization starts from the formal description of a rational agent based on decision theory, according to which agents act so as to maximize the expectation of their performance measure. This decision-theoretic model of decision making can be used to formally define the emotional states of a rational agent. Our definitions capture how emotional states transform the agent's decision-making situation, say, by making the agent more short-sighted, by altering the agent's subjective performance measure (preferences), or by modifying the probabilities that the agent assigns to states of the world for the purpose of expected utility calculations.

Having the formal definitions of emotional states allows us to show how, and why, they are useful. First, the notions of emotional states are useful as factors that allow a rational artificial agent to efficiently control the use of its computational resources [30]. For example, any time pressure that the external environment puts the agent under should lead to it being in states that promote faster, possibly simplified, decision-making. These states, ranging from, say, "rushed" to "panicked", should modify the agent's decision-making situation to exclude weakly relevant information, action alternatives that are likely to be inferior to others, or possibilities that have vanishing probabilities of being realized. Alternatively, emotional states can serve as ready-to-use action tendencies that cut down on the agent's need for deliberative decision-making. For example, if the external environment becomes dangerous, the state of "fear" should trigger an action tendency promoting flight (or an analogous alternative depending on the environment).Second, just as being able to communicate with others about the external environment is useful during interactions, well defined notions of emotional states are valuable when an agent finds it necessary to inform the other agents about its own internal state. Instead of having to describe the details of its internal state, and running the risk of being misunderstood if the other agents are engineered differently, the agent can use more abstract and universal terms. For example, notions of stress or panic are convenient to express the fact that the urgency of the situation forced the agent to look at only short-term effects of its actions. Thus, in the context of communication with other agents, the formal definitions

[0] This research is supported, in part, by the Office of Naval Research grant N00014-95-1-0775, and by the National Science Foundation CAREER award IRI-9702132.